INFORMATION TECHNOLOGY, COMPUTER SCIENCE AND MANAGEMENT
ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

UDC 004.89
https://doi.org/10.23947/2687-1653-2023-23-1-66-75

Original article

Machine Learning Model for Early Detection of COVID-19 by Heart Rhythm Abnormalities

Maksim S Mezhov[1] ✉, **Vyacheslav O Kozitsin**[1], **Iurii D Katser**[2]

[1] "Digital Technologies and Platforms" LLC, 53, Dubininskaya St., Moscow, Russian Federation

[2] Skolkovo Institute of Science and Technology, 30, Bolshoy Boulevard, Moscow, Russian Federation

✉ msmezhov@ya.ru

Abstract

Introduction. Electronic devices capable of collecting individual telemetry data have opened up prospects for preclinical detection of COVID-19 signs. Known solutions involve the analysis of information that is difficult to obtain at the moment. We are talking, specifically, about the blood condition or a PCR test. This significantly limits the possibility of integrating algorithms with wrist gadgets. At the same time, the cardiovascular system as an object of observation is quite informative, the data collection is well developed. The article describes the problem of detecting covid anomalies in rhythm strips. The work aims at creating a mathematical model based on machine learning algorithms to automate the process of detecting covid abnormalities in the heart rhythm. The possibility of integrating the results obtained with fitness bracelets and smart watches is shown.

Materials and Methods. The work involved an open technology stack: Python, Scikit-learn, Lightgbm. When assessing the quality of models for binary classification, metric F_1 was used. 229 cardiac rhythm strips (cardiointervalographies) of patients with COVID-19 were studied. The presence or absence of signs of an anomaly was determined taking into account the time of the rhythm strip and the intervals between heartbeats. Deviations that could indicate infection were shown graphically. Based on the exploratory analysis results, a list of signs indicating an anomaly was made.

Results. As a result of the work done, a mathematical model was obtained that detected heart rate abnormalities specific to COVID-19 with an accuracy of 83 %. The basic features determining the predictive ability of the model were identified and ranked. They included the current value of the interval between heartbeats, the derivatives at the subsequent and previous points of measuring the duration of the heartbeat, the first derivative at the current point, and the deviation of the current value of the duration of the *RR*-interval from the median. The first indicator in this list was recognized as the most significant, the last — the least. For machine learning purposes, the potential of five algorithms was evaluated: IsolationForest, LGBMClassifier, RandomForestClassifier, ExtraTreesClassifier, SGDOneClassSVM. The normal and abnormal results of observations in isolation trees were visualized. A parameter was set that corresponded to the probability of regular observation outside the norm, and its value was selected — 0.11. Taking into account this indicator, a graph was constructed for the SGDOneClassSVM model. Based on the data set, using the cross-validation technique, the quality metric was calculated. The case in hand was a rhythm strip with a time series of observations taken in one continuous time interval from one person. A step-by-step process of obtaining averaged metric values for each model was described. In comparison, the highest indicator was recorded for the LGBMClassifier model, the lowest — for SGDOneClassSVM and IsolationForest.

Discussion and Conclusions. The resulting mathematical model takes up little space in the memory of a mobile device, i.e., it does not impose significant requirements on computing resources. The solution has an acceptable detection quality for pre clinical screening of COVID-19-related cardiovascular disorders. The algorithm detects anomalies in 83 % of cases. Four minutes is enough to record a rhythm strip. The proposed scenario for using an integrated solution is concise and easy to implement. Widespread use of the development can contribute to the detection of COVID-19 at an early stage.

Keywords: COVID-19, causes of death in covid-positive patients, complications in the work of cardiovascular system, PCR test, preclinical monitoring of the cardiovascular system, built-in pulse rate sensors, rhythm strip, RR-interval, cardiac electrocardiogram, abnormal heartbeat, heartbeat with abnormal rhythm, machine learning, LGBMClassifier algorithm.

Acknowledgements. The authors would like to thank the management and moderators of the open All-Russian competition of professionals in the digital economy "Digital Breakthrough" for the data provided for the study.

For citation. Mezhov MS, Kozitsin VO, Katser IuD. Machine Learning Model for Early Detection of COVID-19 by Heart Rhythm Abnormalities. *Advanced Engineering Research (Rostov-on-Don).* 2023;23(1):66–75. https://doi.org/10.23947/2687-1653-2023-23-1-66-75

Научная статья

Модель машинного обучения для обнаружения COVID-19 на ранней стадии по аномалиям в ритме сердца

М.С. Межов[1], В.О. Козицин[1], Ю.Д. Кацер[2]

[1]ООО «Цифровые технологии и платформы», Российская Федерация, Москва, ул. Дубининская, 53, стр. 6

[2]Сколковский институт науки и технологии, Российская Федерация, Москва, территория инновационного центра «Сколково», Большой бульвар, 30, стр. 1

✉ msmezhov@ya.ru

Аннотация

Введение. Электронные устройства, способные собирать данные по телеметрии индивидуума, открыли перспективы доклинического выявления признаков COVID-19. Известные решения предполагают анализ информации, которую сложно получить в моменте. Речь идет, например, о состоянии крови или ПЦР-тесте. Это существенно ограничивает возможности интеграции алгоритмов с наручными гаджетами. При этом сердечно-сосудистая система как объект наблюдения достаточно информативна, съем данных хорошо проработан. В статье описана задача детекции ковидных аномалий в ритмограммах. Цель работы — создание математической модели на базе алгоритмов машинного обучения для автоматизации процесса выявления ковидных аномалий в ритме сердца. Показана возможность интеграции полученных результатов с фитнесс-браслетами и умными часами.

Материалы и методы. В работе задействовали открытый стек технологий: Python, Scikit-learn, Lightgbm. При оценке качества моделей для бинарной классификации использовалась метрика F_1. Изучены 229 ритмограмм сердца (кардиоинтервалографий) пациентов с COVID-19. Наличие или отсутствие признаков аномалии определялось с учетом времени ритмограммы и интервалов между сердцебиениями. Графически показаны отклонения, которые могут свидетельствовать о заражении. По итогам разведочного анализа собран перечень признаков, указывающих на аномалию.

Результаты исследования. В результате проделанной работы получена математическая модель, которая детектирует специфичные для COVID-19 аномалии сердечного ритма с точностью 83 %. Выявлены и ранжированы основные признаки, определяющие прогностическую способность модели. Это текущее значение интервала между ударами сердца, производные в последующей и предыдущей точках измерения продолжительности сердцебиения, первая производная в текущей точке и отклонение от медианы текущего значения длительности *RR*-интервала. Первый показатель в этом перечне признан наиболее значимым, последний — наименее. Для целей машинного обучения оценивался потенциал пяти алгоритмов: IsolationForest, LGBMClassifier, RandomForestClassifier, ExtraTreesClassifier, SGDOneClassSVM. Визуализированы нормальные и аномальные результаты наблюдений в изолирующих деревьях. Установлен параметр, который соответствует вероятности регулярного наблюдения за пределами нормы, и выбрано его значение — 0,11. С учетом данного показателя построен график для модели SGDOneClassSVM. По набору данных с применением техники перекрестной проверки рассчитана метрика качества. Речь идет о ритмограмме с временны́м рядом наблюдений, снятых за один непрерывный интервал времени у одного человека. Описан пошаговый процесс получения усредненных значений метрики для каждой модели. При сравнении самый высокий показатель зафиксирован у модели LGBMClassifier, наименьшие — у SGDOneClassSVM и IsolationForest.

Обсуждение и заключения. Полученная математическая модель занимает мало места в памяти мобильного устройства, то есть не предъявляет значимых требований к вычислительным ресурсам. Решение обладает приемлемым качеством детекции для доклинического скрининга связанных с COVID-19 сердечно-сосудистых

нарушений. Алгоритм обнаруживает аномалии в 83 % случаев. Для записи ритмограммы достаточно 4 минут. Предлагаемый сценарий использования интегрированного решения лаконичен и легко реализуем. Широкое использование разработки может способствовать выявлению COVID-19 на ранней стадии.

Ключевые слова: COVID-19, причины смерти ковид-положительных пациентов, осложнения в работе сердечно-сосудистой системы, ПЦР-тест, доклинический контроль сердечно-сосудистой системы, встроенные датчики частоты пульса, ритмограмма, RR-интервал, электрокардиограмма сердца, аномальное по продолжительности сердцебиение, сердцебиение с аномальным ритмом, машинное обучение, алгоритм LGBMClassifier.

Благодарности. Авторы выражают благодарность руководству и модераторам открытого всероссийского соревнования профессионалов в сфере цифровой экономики «Цифровой прорыв» за предоставленные данные для исследования.

Для цитирования. Межов М.С., Козицин В.О., Кацер Ю.Д. Модель машинного обучения для обнаружения COVID-19 на ранней стадии по аномалиям в ритме сердца. *Advanced Engineering Research (Rostov-on-Don).* 2023;23(1):66–75. https://doi.org/10.23947/2687-1653-2023-23-1-66-75

Introduction. Investigation of the impact of COVID-19 on humans remains a challenge. Thus, in 2021–2022, more than 16,000 scientific papers were published on this topic. One of the main causes of death of covid-positive patients was complications in the cardiovascular system (hereinafter referred to as CVS) caused by exposure to coronavirus [1]. Two methods are mainly used for preclinical diagnosis of COVID-19: biochemical method based on polymerase chain reaction (PCR test) and blood analysis. Contacts with medical staff needed in this case (including visits to medical institutions) complicate regular operational control and increase the burden on the healthcare system. Thus, it seems relevant to use modern technologies of preclinical control of CVS for early detection of COVID-19 signs.

Wearable electronic devices can provide regular monitoring. The most common of them are fitness bracelets and smart watches with built-in heart rate sensors and the ability to perform measurements with high discreteness [2]. This approach opens up opportunities for analyzing data flows based on machine learning[1] [3].

The presented study aims at creating a trainable model capable of detecting covid anomalies based only on data on heart rhythm. A number of papers [4–6] consider similar problems, but the solutions are based on additional information about the state of the blood and other characteristics[2]. This significantly limits the possibilities of their integration with wearable devices, because at the moment, it is impossible to enter the results of a blood test or a smear for a PCR test into the model. The novelty of the proposed solution is in the fact that only heart rate data is used, which can be taken with a high frequency in a way convenient for a person and interpret the indicators in real time.

Materials and Methods

Data characteristics. 229 impersonal rhythm strips (cardiointervalographies) of patients with COVID-19 were used in the research. The information was obtained in 2021 as part of the open All-Russian competition "Digital Breakthrough" for professionals in the digital economy. A data fragment is presented in Table 1.

Table 1

A fragment of the data set

Number of rhythm strip	Time in milliseconds	RR interval between heartbeats in milliseconds	Sign of covid anomaly*
81	0	576	0
81	568	568	0
81	1,140	572	0
…	…	…	…
176	44,332	568	0
176	44,968	636	1
176	45,596	628	0
*0 — no anomaly, 1 - there is an anomaly.			

Figure 1 shows the relationship of the rhythm strip (RR interval) and the electrocardiogram of the heart (ECG).

[1] Permyakov SA, et al. Ehndogennye anomalii kardioritma u patsientov s COVID-19. In: Proc. VII All-Russian Conf. "Nelineinaya dinamika v kognitivnykh issledovaniyakh – 2021". Nizhny Novgorod: Institute of Applied Physics of RAS; 2021. P. 109–110. (In Russ.)

[2] Diagnosis of COVID-19 and Its Clinical Spectrum. Kaggle Inc. URL: https://www.kaggle.com/datasets/einsteindata4u/covid19 (accessed: 10.09.2022).

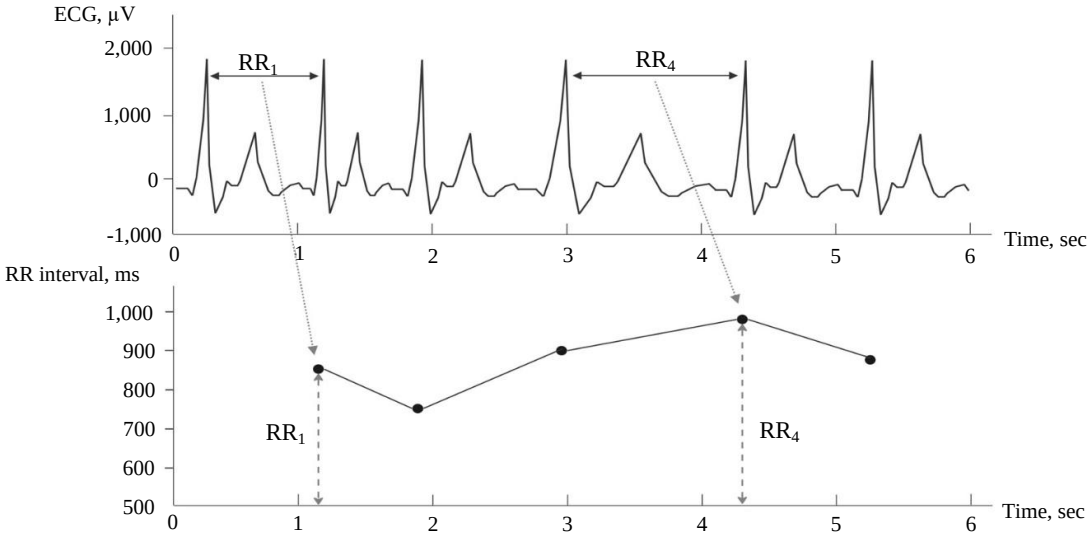

Fig. 1. Comparison of electrocardiogram and heart rhythm strip: horizontal axis shows the time in seconds, vertical axis for the ECG — microvolts

In all rhythm strips from this set, there are marked abnormal areas. In Figure 2, abnormal areas are highlighted with a red dotted line. The x-axis shows the duration of one measurement of the rhythm strip in milliseconds, the y-axis — the interval between adjacent heartbeats in milliseconds.

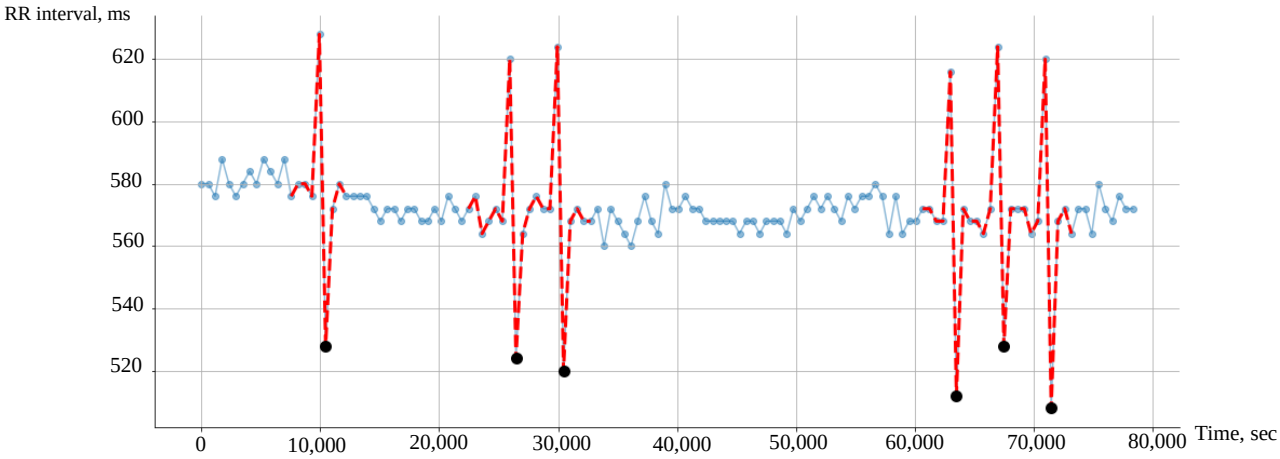

Fig. 2. Chart of rhythm strip No. 69: abnormal areas are highlighted with red dotted lines, abnormal points — with black bullet points

Each rhythm strip is presented by its own identifier. The duration of rhythm strips in the studied data set is different: 4 minutes on average, 31 minutes maximum. Each measurement inside one rhythm strip has a timestamp in milliseconds from the start of recording. The duration of the RR interval is also presented in milliseconds. Each specific value in the rhythm strip allows us to talk about the signs of an anomaly (0 — no, 1 — there is). 2.53 % of observations are marked with number 1. Thus, the data set has a strong class imbalance, which is typical for anomaly detection tasks.

In the data markup, there are various approaches to the allocation of abnormal areas. Groups of points in the vicinity of a characteristic peak and fall in the duration of the heart rhythm were distinguished as abnormal: 3rd, 4th, 6th measurements (Fig. 2). Not always the number of points in the neighborhood is marked the same — there may be a different number of abnormal points to the left and right of the peak. Moreover, rhythm strips with noisy indications were detected. This was the case when the connection with the gadget was lost, and measurements were taken when installing or removing the device. Sixteen rhythm strips with incorrect data had to be excluded from consideration, and the markup was redone:

– only one point stands out in the anomalous section, characterizing the anomalous fragment;

– abnormal points are indicated by black bullets (Fig. 2).

Feature extraction. In its pure form, only one signal was presented — the value of intervals between heartbeats. Therefore, to refine the model, additional features were prepared based on the available signal: deviation from the median value and derivatives in neighboring rhythm measurements. This list of features was selected after an

exploratory data analysis and visual identification of the pattern in places corresponding to abnormal areas. In Figure 2, they were marked with a red dotted line.

Research Results

Metric for evaluating the quality of anomaly detection. To assess the quality of the model in the binary classification problem, due to the imbalance of classes, metric F_1 [7] (1) was used. It provided evaluating how well a constructed model detected a rare class. In that context, a rare class referred to abnormal heartbeats in duration — heartbeats with an abnormal rhythm:

$$F_1 = 2 \times \frac{\text{accuracy} \times \text{completeness}}{(\text{accuracy} + \text{completeness})}. \quad (1)$$

Here:

– accuracy — the proportion of abnormal heartbeats correctly detected by the model from the total number of heartbeats that the model identified as abnormal;

– completeness (or in other words, sensitivity) — the proportion of heartbeats that the model correctly detected as abnormal from the total number of abnormal heartbeats in the entire data set.

Machine learning algorithms. As part of the study, five machine learning algorithms described below were applied.

1. IsolationForest — an algorithm with uncontrolled self-learning based on extremely randomized decision trees [8].

2. Light Gradient Boosting Machine Classifier (LGBMClassifier) — an algorithm for gradient boosting over decision trees [9]. To increase the operation speed, two techniques were used: Gradient-based One-Side Sampling and Exclusive Feature Bundling[3].

3. RandomForestClassifier is based on decision trees and implements multiple selection of a random subset of features. They are used to build simpler estimators — decision trees. The results are aggregated to obtain a final prediction [10].

4. ExtraTreesClassifier is similar to RandomForestClassifier, however, it additionally implements a random selection of the boundary along which nodes branch in decision trees [11].

5. SGDOneClassSVM[4] — a linear version of One-Class Support Vector Machine using Stochastic Gradient Descent.

IsolationForest and SGDOneClassSVM were chosen due to their wide use in anomaly detection tasks [12, 13]. LGBMClassifier, RandomForestClassifier and ExtraTreesClassifier perform well enough in different tasks, therefore, they were also used to compare the results [12, 13].

The specific feature of the IsolationForest and SGDOneClassSVM algorithms is that they do not require a clear marking of anomalous observations at the input, while it is mandatory for the rest of the algorithms used in the study.

IsolationForest is based on the assumption that when constructing isolating trees, abnormal observations can be isolated (separated) in fewer operations than normal observation instances. For each observation, the algorithm calculates the anomaly score by the formula:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (2)$$

where $h(x)$ — number of edges up to instance x in each isolating decision tree; $E(h(x))$ — average value $h(x)$ on the entire set of isolating trees; $c(n)$ — normalizing constant for a data set of size n (3).

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \quad (3)$$

$$H(k) = \ln(k) + \gamma. \quad (4)$$

In equation (4) γ — Euler's constant equal to 0.57721…

If observation x has an anomaly estimation value s, close to 1, then it is considered anomalous. If s is close to 0.5, then the observation has no obvious signs of an anomaly. If s is close to 0, then the observation can be considered normal (Fig. 3).

[3] LightGBM: A Highly Efficient Gradient Boosting Decision Tree. www.microsoft.com URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/lightgbm.pdf (accessed: 10.09.2022).

[4] Online One-Class SVM. Scikit-learn developers (BSD License). scikit-learn.org URL: https://scikit-learn.org/stable/modules/sgd.html#online-one-class-svm (accessed: 10.09.2022).

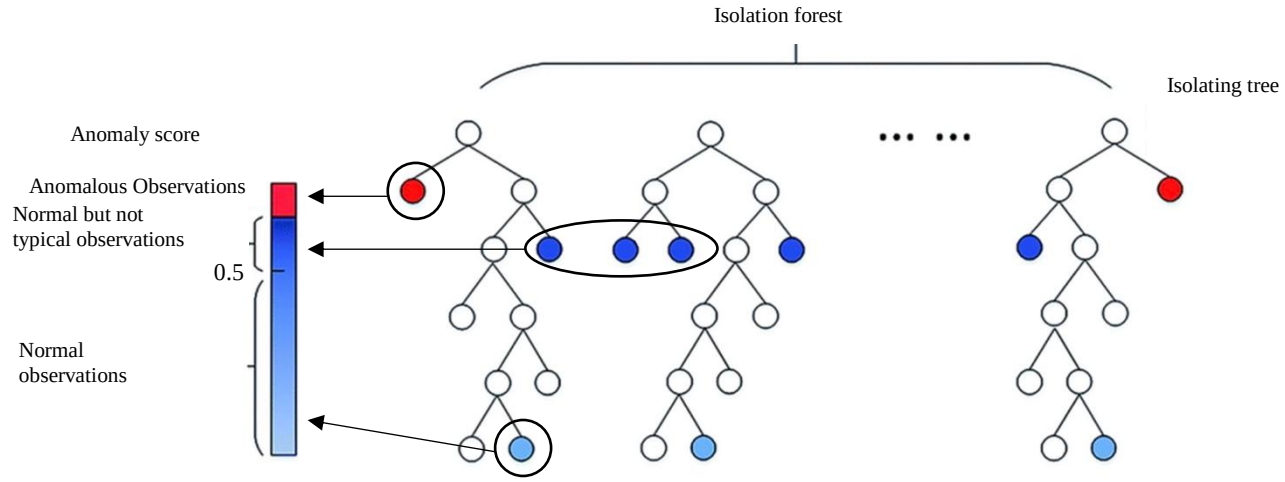

Fig. 3. Normal and anomalous observations in isolating trees (the authors' figure)

SGDOneClassSVM is based on the opposite approach to IsolationForest. The algorithm determines the boundaries of normal observations and compares all new observations to the boundaries of this norm to identify an anomaly.

Feature Significance. An assessment of the degree of impact of features on the predictive ability of the model is shown in Figure 4.

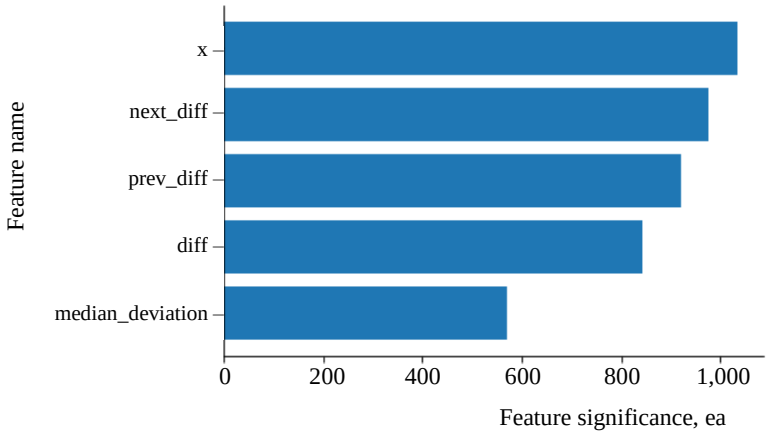

Fig. 4. Feature significance diagram: *x* — current value of the interval; next_diff — derivative at the next point of measuring the heartbeat duration; prev_diff — derivative at the previous point of measuring the heartbeat duration; diff — the first derivative at the current point; median_deviation — deviation of the current value of the RR interval duration from the median

To calculate the numerical significance estimate, a mechanism built into LGBMClassifier was used, which returns an array of numerical estimates for each feature via the *feature_importances_* property of the trained model. Significance in models based on gradient boosting over decision trees is usually calculated on the Gini-impurity Index[5] [14], used in the process of determining the branching points when training the model:

$$Gini(d) = 1 - \sum_{i=1}^{k} p_i^2 \,. \quad (5)$$

Here, d — a set of observations that match the conditions at the considered branching point, $d \in D$; k — number of classes presented in the entire training dataset D; p_i — probability of observations belonging to class i at the considered branching point of the decision tree.

The following features were the most significant: the current value of interval (x), the derivative at the next (next_diff) and previous (prev_diff) points of measuring the heartbeat duration (Fig. 4). A complete list of the features used is given in Table 2.

[5] Karabiber F. Gini Impurity. learndatasci.com URL: https://www.learndatasci.com/glossary/gini-impurity/ (accessed: 10.09.2022).

Table 2

List of features used

No.	Feature	Description
1	*x*	RR interval at the current measuring point
2	next_diff	First derivative at the next point
3	prev_diff	First derivative at the previous point
4	diff	First derivative at the current point
5	median_deviation	Deviation of the current value of the RR interval duration from the median within one rhythm strip

Comparison of models. For effectiveness of SGDOneClassSVMmodel, it is important to select parameter *nu*, which corresponds to the probability of detecting regular observation outside the norm. In other words, *nu* determines the upper bound of the error rate when training the model, and the lower bound of the support vector fraction[6]. To select *nu* taking into account the available data nature, the quality metric was additionally assessed at different values of the specified parameter (Fig. 5). As a result, *nu* equal to 0.11 was selected.

.

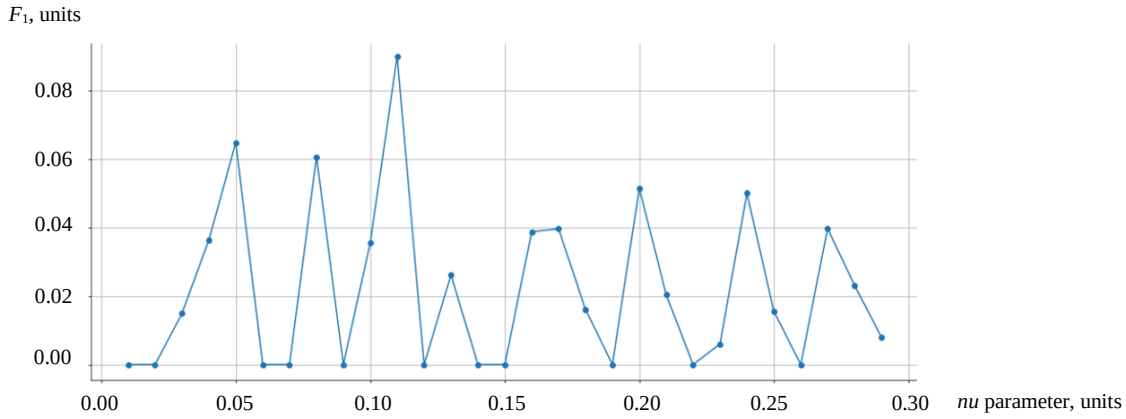

Fig. 5. Estimate of parameter *nu* (on the horizontal axis) for *SGDOneClassSVM* model.
On the vertical axis — values of metric F_1

To calculate the quality metric on various models, the entire data set was used through the cross-validation technique. Within one rhythm strip, we had a time series of observations taken in one continuous period of time from one person, therefore, they should be considered as dependent [15]. The following strategy was used to divide the data into training and test sets. The selected data set consisted of 213 rhythm strips marked with a unique identifier (*id*). This made it possible to allocate rhythm strips for training and testing models. A set of rhythm strips for the test could be randomly selected by identifiers. The approach used in the presented work is described below.

I. Five actions were performed in the data partitioning cycle.

1. The initial number for generating pseudo-random numbers was fixed (seed) —np.random.seed(fold), where fold — number of the current data partition.

2. 42 random integer values were generated in the range from 1 to 213. This was how we got random numbers of rhythm strip identifiers for the test data set.

3. The numbers of rhythm strip identifiers that remained after the selection of identifiers for the test were entered in a separate list. They were used for a training set.

4. Models were trained on rhythm strips from the training set, and prediction quality metrics were evaluated on rhythm strips from the test set.

5. The quality metric value was recorded for each model calculated on the test set of rhythm strips at the current data split.

[6] SGDOneClassSVM documentation. Scikit-learn developers (BSD License). scikit-learn.org URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDOneClassSVM.html#sklearn.linear_model.SGDOneClassSVM (accessed: 10.09.2022).

II. Steps 1–5 were repeated for each data split number.

III. The obtained values of the quality metric were averaged for each of the models.

A comparative assessment of the average prediction quality metric for each model is given in Table 3.

Table 3

Evaluation of quality metric F_1

Model	Metric F_1*
LGBMClassifier	0.8328
RandomForestClassifier	0.7638
ExtraTreesClassifier	0.7369
SGDOneClassSVM	0.0169
IsolationForest	< 1e-4
* Average value for the selected cross-validation strategy on five partitions.	

Discussion and Conclusions. A mathematical model for detecting anomalies in the heart rhythm with an accuracy of 83 % has been developed. According to quality metric F_1, the model based on LGBMClassifier algorithm turned out to be the best. IsolationForest and SGDOneClassSVM showed weak results on current data.

The proposed model can be implemented as a component of the software part of wearable personal smart devices. The proposed scenario for using the solution is as follows:

– the recording of the rhythm strip is activated on a personal wearable device through the user interface;

– upon completion, the record is submitted to the developed model for analysis;

– based on the results of data analysis, the mathematical model issues a notification about the presence or absence of anomalies on the screen of the wearable device.

Note that an average of 4 minutes is probably enough to record one rhythm strip. During this time, it is possible to detect covid anomalies in the heart rhythm.

The model occupies 493 kilobytes in the memory of the wearable device, which is quite suitable for practical use. The solution relies only on information about the heart rate and does not involve factors inaccessible to mobile personal gadgets.

Improving the accuracy of anomaly detection involves additional research. They should focus on the development of unique features that are detected by the initial heart rate signal. However, the current solution already makes it possible to quickly and easily assess the probability of COVID-19 at an early stage. This, along with the implementation of medical recommendations, can further contribute to reducing the risk of mortality from the negative impact of coronavirus infection on the cardiovascular system.

References

1. Tursunova ND, Shafigulina IS, Grebennikova IV, et al. Patogeneticheskie aspekty vliyaniya COVID-19 na serdechno-sosudistuyu sistemu cheloveka. *European Journal of Natural History.* 2022;1:73–77. (In Russ.)

2. Molodchenkov AI, Grigoriev OG, Sharafutdinov YaN. Automatic Calculation of Disease Risk Factors Values Using Artificial Intelligence Methods and Internet of Things Technology. *Journal of Information Technologies and Computing Systems.* 2021;1:83–96. https://doi.org/10.14357/20718632210109

3. Polevaya SA, Eremin EV, Bulanov NA, et al. Event-Related Telemetry of Heart Rhythm for Personalized Remote Monitoring of Cognitive Functions and Stress under Conditions of Everyday Activity. *Modern Technologies in Medicine*. 2019;11:109–115. http://dx.doi.org/10.17691/stm2019.11.1.13

4. Kouame Amos Brou, Ivan Smirnov, Mabouh Moise Hermann. Comparison of Machine Learning Models for Coronavirus Prediction. *Advanced Engineering Research (Russia).* 2022;22:67–75. https://doi.org/10.23947/2687-1653-2022-22-1-67-75

5. Ashish Bhargava, Elisa Akagi Fukushima, Miriam Levine, et al. Predictors for Severe COVID-19 Infection. *Clinical Infectious Diseases.* 2020;71:1962–1968. https://doi.org/10.1093/cid/ciaa674

6. Krasyukova YuI, Vakhrusheva TA, Pei He Su. Machine Learning Model for Determining the Probability of Covid-19 Disease by Primary Signs. *Intellektual'nye resursy — regional'nomu razvitiyu*. 2021;2:67–71.

7. Alaa Tharwat. Classification Assessment Methods. *Applied Computing and Informatics.* 2021;17:174. https://doi.org/10.1016/j.aci.2018.08.003

8. Yupeng Xu, Hao Dong, Mingzhu Zhou, et al. Improved Isolation Forest Algorithm for Anomaly Test Data Detection. *Journal of Computer and Communications.* 2021;9:49–51. https://doi.org/10.4236/jcc.2021.98004

9. Bruce P, Bruce A, Gedeck P. *Practical Statistics for Data Scientists,* 2nd ed. Boston: O'Reilly Media; 2020. 342 p.

10. Breiman L. Random Forests. *Machine Learning.* 2001;45:5–32. https://doi.org/10.1023/A:1010933404324

11. Geurts P, Ernst D, Wehenkel L Extremely Randomized Trees. *Machine Learning.* 2006;63:3–42. https://doi.org/10.1007/s10994-006-6226-1

12. Kaur H, Singh G, Minhas J. A Review of Machine Learning Based Anomaly Detection Techniques. *International Journal of Computer Applications Technology and Research.* 2013;2:185–187. http://dx.doi.org/10.7753/IJCATR0202.1020

13. Katser ID, Kozitsin VO, Maksimov IV. NPP Equipment Fault Detection Methods. *Proc. of Universities. Nuclear Power Engineering*. 2019;4:5–27. https://doi.org/10.26583/npe.2019.4.01

14. Daniya T, Geetha M, Suresh Kumar K Dr. Classification and Regression Trees with Gini Index. *Advances in Mathematics Scientific Journal*. 2020;9:8237–8247. http://dx.doi.org/10.37418/amsj.9.10.53

15. Valliappa Lakshmanan, Sara Robinson, Michael Munn. *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps,* 1st ed. Boston: O'Reilly Media; 2020. 408 p.

About the Authors:

Maksim S Mezhov, leading expert, "Digital Technologies and Platforms" LLC (53, Dubininskaya St., Moscow, 115054, RF), ORCID, msmezhov@ya.ru

Vyacheslav O Kozitsin, leading expert, "Digital Technologies and Platforms" LLC (53, Dubininskaya St., Moscow, 115054, RF), ORCID, Vyacheslav.Kozitsin@skoltech.ru

Iurii D Katser, postgraduate, Skolkovo Institute of Science and Technology (30, Bolshoy Boulevard, Moscow, 121205, RF), ScopusID, ORCID, Iurii.katser@skoltech.ru

Claimed contributorship:

MS Mezhov: basic concept formulation; research objectives and tasks; data collection; model development; calculations and analysis of the results. VO Kozitsin: text preparation; formulation of conclusions; pre-processing of data; the text revision. IuD Katser: control of the study; revision of the text; correction of the conclusions.

Received 09.12.2022.
Revised 25.01.2023.
Accepted 25.01.2023.

Conflict of interest statement

The authors do not have any conflict of interest.

All authors have read and approved the final manuscript.

Об авторах:

Максим Сергеевич Межов, ведущий эксперт ООО «Цифровые технологии и платформы» (115054, РФ, Москва, ул. Дубининская, 53, стр. 6), ORCID, msmezhov@ya.ru

Вячеслав Олегович Козицин, ведущий эксперт ООО «Цифровые технологии и платформы» (115054, РФ, Москва, ул. Дубининская, 53, стр. 6), ORCID, Vyacheslav.Kozitsin@skoltech.ru

Юрий Дмитриевич Кацер, аспирант сколковского института науки и технологии (121205, РФ, Москва, территория инновационного центра «Сколково», Большой бульвар, 30, стр. 1), ScopusID, ORCID, Iurii.katser@skoltech.ru

Заявленный вклад соавторов

М.С. Межов — формирование основной концепции, цели и задач исследования, сбор данных, разработка моделей, расчеты и анализ результатов. В.О. Козицин — подготовка текста, формулирование выводов, предварительная обработка данных и доработка текста. Ю.Д. Кацер — контроль проведения исследования, доработка текста и корректировка выводов.

Поступила в редакцию 09.12.2022.
Поступила после рецензирования 25.01.2023.
Принята к публикации 25.01.2023.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Все авторы прочитали и одобрили окончательный вариант рукописи.